

Ternary training and post-training quantization

Trit3, 15 October 2026

We compared ternary training with projection-only post-training quantization of float models across two small model sizes. At the largest training-data point, ternary models outperformed the best tested 2-bit quantization, while the best 3- and 4-bit alternatives had lower loss. Interpolated quality-equivalent nominal precision was approximately 2.89 bits for size T and 2.82 bits for size S. These are local quality comparisons; they do not establish equal-total-memory product performance.

The question

How does ternary training compare with compressing a float model after training? We evaluated projection-only round-to-nearest (RTN) and GPTQ quantization at 2, 3 and 4 bits, using the 36 available W3 models: two T seeds and one S seed, six training-data points each, with float and ternary counterparts.

New evaluations used L4 hardware and FP32 computation with the unchanged W3 evaluator. Ternary losses used the W3 document arrays after hardware checks. The GPTQ implementation was corrected prospectively to match the author-reference grid and update rules; the earlier implementation remains a sensitivity analysis.

Quality at the largest training-data point

Lower negative log-likelihood (NLL) is better. The best quantization method was selected separately at each bit width within each bootstrap replicate.

Model / method	Size T NLL	Size S NLL
Float	4.8900	4.7202
Ternary	5.0897	4.8866
Best 4-bit PTQ	4.9124	4.7349
Best 3-bit PTQ	5.0038	4.7918
Best 2-bit PTQ	5.7799	5.3095

Quality-equivalent nominal precision was 2.8893 bits (95% CI: 2.8860-2.8927) for T and 2.8170 bits (2.8122-2.8216) for S. This interpolation uses adjacent 2-, 3- and 4-bit points only. It is not a measured storage budget: nominal bit widths exclude packing and scale overhead.

Training data and the tail

The degradation from both 3- and 4-bit RTN and GPTQ increased with training data in both sizes over the frozen fitting range. This replicates the tested direction at this scale; it does not establish a universal law.

Tail concentration depended on the method and size. For T, GPTQ 4-bit and ternary training concentrated more of their loss gap in the union tail, while RTN 4-bit did not. For S, RTN 4-bit and ternary training showed concentration; GPTQ 4-bit was inconclusive. The union mask combines the study's frequency and token-category proxies without counting overlapping categories twice.

A separate BitNet preparation check

An implementation using the pinned official BitNet architecture matched the official BF16 forward computation on a fixed reference sample. A short runtime benchmark identified an A10, rank-16 configuration as the cheapest measured setting in that test. This prepares a future intervention study; it is not a quality improvement result or a verification of bitnet.cpp kernel parity.

What this result does and does not establish

The models are small and use context length 256. Embeddings and the output head dominate whole-model memory, so projection-level comparisons cannot be read as equal-total-memory product comparisons. S has one training seed. The 95% intervals use 1,000 paired whole-document bootstrap replicates and condition on fixed trained models; they do not estimate training-seed population variability.

Category correspondence does not establish a shared causal mechanism. No correction to model quality was tested here. The next experiment is a separately designed test of a specific intervention at a larger scale.